DOCUMENT RESUME

ED 397 092 TM 025 130

AUTHOR Minke, Amy
TITLE A Review of the Recent Changes in the Scholastic Aptitude Test I: Reasoning Test.
PUB DATE 26 Jan 96
NOTE 14p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (New Orleans, LA, January 1996).
PUB TYPE Information Analyses (070) -- Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Change; College Bound Students; *College Entrance Examinations; High Schools; High School Students; Item Response Theory; *Mathematics Tests; National Norms; Predictive Validity; *Scoring; Test Construction; Test Content; Test Reliability; *Test Use; Test Validity; *Thinking Skills; *Verbal Tests
IDENTIFIERS *Scholastic Aptitude Test; Scholastic Assessment Tests; Test Specifications

ABSTRACT

The Scholastic Aptitude Test (SAT) has been designed to test developed verbal and mathematics reasoning abilities of college-bound students, primarily high school juniors and seniors. For almost a decade there has been a research and development process to evaluate and change the entire SAT program. These changes were implemented in the SAT I: Reasoning Test in content and administration changes in April 1994 and scale recentering in April 1995. As of October 1995, a technical manual for the SAT I had not been published, but extensive research on the SAT I has been reported. The SAT I is normed on the 1990 reference group of 1,052,000 scores from 35 editions from October 1988 through June 1990. Field trials of the new test involved 162,692 high school juniors. By modeling the new test using item response theory, it has been estimated that the reliability of the new test is comparable to the old one, if not better. Definitive studies have not been done on test validity but research has suggested that recentering the scale has improved its predictive validity. Pains were taken in the test improvement process to ensure that the new test was not easier than the old one, even though mean test scores "rose." It could be said that changes to the SAT were not only justified but sorely needed to serve the test taking population adequately. Research on the effects of the new test specifications is needed, especially for changes in the mathematics section, that occurred after the field studies. (Contains 18 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
AMY MINKE

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

ED 397 092

A Review of the Recent Changes in the Scholastic Aptitude Test I: Reasoning Test

Amy Minke
Graduate Student
Department of Educational Psychology
Texas A&M University

Paper presented at the annual meeting of the
Southwest Educational Research Association
New Orleans, Louisiana
January 26, 1996

BEST COPY AVAILABLE

TM025130

General Information

Test: Scholastic Aptitude Test I: Reasoning Test (SAT I)
Publisher: Educational Testing Service and the College Entrance Examination Board
Date of Publication: 1994 & 1995
Time Required: Three hours
Cost: $21.50

The SAT I is a group administered test. The time required and the cost reported above are for individuals. Each individual taking the test registers independently. The cost covers test materials, basic registration and administration costs, a personal score report, and four score reports to colleges or scholarship programs.

Background and History

The Scholastic Aptitude Test, more commonly known as the SAT, has been designed to test developed verbal and mathematical reasoning abilities of college-bound students, primarily high school juniors and seniors. Since these abilities are related to successful performance in college, results from the SAT are used to better assess the readiness of college applicants for college work. The scores are intended to be used in conjunction with secondary school records and other information about the aptitude and motivation of applicants in order to make a more informed admissions decision.

The SAT became a part of the College Board in 1926 in an effort to standardize college admissions testing and has been known for its exemplary development and technical merits (Anastasi, 1988). It has undergone changes since its development, but for almost a decade there has been a research and development process to evaluate and change the entire SAT program. These changes were implemented in the SAT I: Reasoning Test in two phases: content and administration change in April 1994 and scale recentering in April 1995 (Cook, 1995).

Changes

The time had come for change. Since 1941, the SAT was normed on just over 10,000 test takers from the Spring of 1941. These test takers represented a relatively small and homogeneous college-bound population. Since then, the SAT test-taking population, just as the college-bound population, has become much larger and more diverse, exceeding 1,000,000 test takers in a year (Dorans, 1994a).

This demographic change in test takers has resulted in the mean scores sliding down from 500, the mid-point of the characteristic 200-800 standard score SAT scale, to 422 for the Verbal section and

475 for the Math section for the 1990 Reference Group (Dorans, 1994b). This slide is due, at least in part, to more individuals of lesser ability being tested than was the case in 1941. The unequal nature of the two section means caused confusion in interpretation. For example, a 450 on both sections did not represent equal ability in mathematics and verbal reasoning, but above average ability on the Verbal and below average ability on the Math. The skewed nature of the means also meant that there was not a one-to-one correspondence between items and points gained at every scale level. The lower part of the scale was compressed with scores below 200 rounded up to 200, while the upper part of the scale was spread with scores being extended to reach the upper limit of 800. In other words, if a low scoring individual answered one more item correctly, the score could jump less than 10 points, while if a high scoring individual answered one more item correctly, the score could jump by more than 10 points, sometimes as much as 60 points at the high end of the Verbal scale (Dorans, 1994a).

The SAT Program also recognized a need to keep abreast of new theory in cognition, learning, and psychometrics, as well as national curriculum standards set by organizations such as the National Council of Teachers of Mathematics, National Council of Mathematics, Mathematics Sciences Education Board, Mathematical Association of America, and American Mathematical Society. The SAT Program deemed that change was necessary and beneficial to both test takers and users in order to better test and represent the current college bound population's reasoning abilities (College Entrance Examination Board, 1991).

Discussion on how best to change the SAT was begun in 1986 with the New Possibilities Project, including focus group discussions with students, parents, high school guidance counselors, and college admissions officers. Change was not taken lightly since the SAT affects the lives and workings of many individuals. Through this process, it was decided to change the content, administration, and scaling of the SAT in several ways. These changes necessitated a new name, the SAT I: Reasoning Test.

The test consisted of two 30-minute timed sections for both the Verbal and Mathematical tests. In addition to these four sections, there was another Verbal or Mathematical section included for the purposes of equating and pretesting of items, as well as a 30-minute section of the Test of Standard Written English (TSWE). The Test of Standard Written English (TSWE) was intended to be used for placement in freshman English classes, but was being used by few colleges and universities. Thus, it was

removed, allowing an additional 15-minute timed section to be added for both the Verbal and Mathematical tests (Cook, 1995).

The Verbal test was made up of 85 items of four multiple choice types: reading comprehension, analogies, sentence completion and antonyms. In order to keep up with current theories of learning which emphasize contextual knowledge, the antonym items were removed and reading comprehension items were replaced by critical reading items. Now, passage-based items represent 50% of the Verbal items, up from 26%. Also, the reading passages are longer to accommodate the shift in emphasis from word knowledge to contextual reading abilities; because the passages required more testing time, this necessitated a shorter test of 78 items (Cook, 1995). The Math section was made up of 60 multiple choice items of two types: 40 five-choice mathematical concept items and 20 four-choice quantitative comparison items. The total number of items remained the same, but the new test includes 35 five-choice mathematical concept items, 15 four-choice quantitative comparison items, and 10 Student Produced Response (SPR) items. Also, test takers are encouraged to use calculators on the Math sections, a change to keep abreast of current classroom practice. The topics of the items now better fit curriculum changes, focusing on applying mathematical concepts and interpreting data. These content changes were introduced for the tests administered beginning in April 1994.

In April 1995, the SAT scale was recentered on a more current, representative group , the 1990 Reference group. The original tests had means of 500, the mid-point of the 200-800 score reporting range, and standard deviations of 100, allowing for three standard deviations from the mean to be covered by the scale. The means for both sections were reestablished at 500, but with a standard deviation of 110 (Dorans, 1994a).

Technical Considerations

As of October 1995, a technical manual for the SAT I had not been published. However, the staffs of Educational Testing Service, the test administrators, and the College Entrance Examination Board, the test producers, have conducted extensive research on the SAT I. The studies have been reported as research papers. Relevant papers are referenced in this paper, indicating their salient points. One paper, not mentioned specifically in the text of this paper, is Dodd (1993) which consists of discussant comments from a symposium presentation of Feryok and Wright (1993), Lawrence and Schmitt (1993), and Dorans and Feigenbaum (1993).

The SAT I is normed on the 1990 Reference Group which is made up of the most recent scores of high school students who graduated in 1990 and who last took the SAT in their junior or senior year. The 1990 Reference Group is different from the 1990 College-Bound Seniors Cohort who graduated in 1990 and last took the SAT any time in high school through March of their senior year. The 1990 Reference Group is made up of 1,052,000 scores from 35 editions of the SAT administered between October 1988 and June 1990. The SAT test taking population is relatively invariant now from a demographic point of view, so this group is believed to represent adequately the test taking population (Dorans, 1994a). It is important to note that the norm group contains only college-bound juniors and seniors because that is the group that the test is designed to measure and that should have the most motivation to perform well on the test.

With the content changes, came statistical specification changes in the test. These affect the item difficulty distribution, as an effort was made to keep the average difficulty the same from the old to the new test. Lawrence and Schmitt (1993), discuss these changes in great detail. The changes occurred in two phases, first for the field study and then revisions for the new test. The field trials took place in the spring of 1992 and consisted of 162,692 high school juniors from 2,221 volunteer school districts. Efforts were made to ensure that all test taking subgroups of the population were adequately represented in the sample in order to provide useful results (Feryok and Wright, 1993). The new test specifications for the field trials were modeled using Item Response Theory, keeping the average difficulty of the test the same and the Standard Error of Measurement relatively constant across all levels of the scale. With the field trial results, further specification changes were made for the new test.

Changes in specifications for the Verbal test resulted from the change in item types. The antonym items on the old test constituted the majority of the difficult items on the test. Also, the shift in emphasis to contextual reading resulted in longer reading passages, requiring more time and consequently less difficult items. Most of the available critical reading items are of moderate difficulty, which limited the choice of difficult items. The field trial results matched the expected results modeled from the specifications and few changes were made to the Verbal test from the field trial to the new test.

In the Math section, the specifications changed because of two factors. The Student Produced Response (SPR) items provided a challenge. There was no existing data on them, so for the field studies the test was built using items that modeled SPR item qualities, which are more difficult because of the absence of a guessing factor and the longer time required to create answers. The use of calculators provided another challenge. For the field trials, no changes in statistical specifications were made for the calculator use in order to determine it's affects. The field trial results indicated that the use of calculators made some items easier. Consequently, the test was easier and more difficult items would need to be used in order to offset the calculator use. The results also provided actual data on the SPR items. Researchers determined that the item difficulty scale being used was inadequate to deal with the mix of items with and without guessing factors. They expressed Mathematical items on a new difficulty scale which differentiates between the multiple choice items with guessing factors and the SPR items with no guessing factors. With the new difficulty scale, changes in the specifications have been made for the new test (Lawrence and Schmitt, 1993). There is no reported research for the Mathematics test built to the new specifications.

By modeling the new test using Item Response Theory, it was estimated that the reliability of the new test is comparable, if not better than the old one. Three types of reliability coefficients were compared: the Dressel adaptation of the KR-20, used for formula scored tests; and two estimates of total test reliability, the Angoff-Feldt method, used for two timed sections; and the Kristof method, used for three timed sections. For the Verbal test, the KR-20 coefficients for the old and prototype tests were .920 and .921, respectively. The Angoff-Feldt coefficient was .910 for both versions; for the prototype test, the 15-minute section was grouped with one of the 30-minute sections. This compared to the Kristof coefficient for the prototype test of .908 (Lawrence and Schmitt, 1993). Because there are more items related to each reading passage, the Verbal test reliability could have been affected by increased item dependence (Lawrence 1995), as well as a more unimodal distribution of item difficulty.

Lawrence (1995) presents reliability information for the first seven administrations of the new Verbal test, before recentering. The variance component total test reliability coefficients, based on the Dressel KR-20 coefficients for each timed section, were .922 for the old test and .925 for the new test. The Angoff-Feldt coefficients were .915 for both the old and new tests, while the Kristof coefficient was .913 for the new test. A parallel forms reliability coefficient was computed using test takers who

BEST COPY AVAILABLE

took the test within one or two test administrations. The coefficient for the old test was .898 while the coefficient for the new test is .888. In general, the results for the new test are slightly less reliable. This is mostly attributable to having fewer items, but may be due in part to larger item sets with common stimulus material.

For the Mathematics test, the SPR items appear to be more reliable. Unpublished research referred to by Cook (1995) states that the analysis from the first seven administrations after the new content changes yielded reliability for the SPR item set ranging from .73 to .81. Reliability coefficients based on test specifications for the field trial showed old and new KR-20 coefficients of .925 and .919, respectively. For the old test the Angoff-Feldt coefficient of .922 compared to .912 when the 15-minute section was grouped with the first 30-minute section, and .907 when the 15-minute section was grouped with the second 30-minute section The Kristof coefficient for the new prototype was .909. No data were available for the new specifications of the Mathematics test (Lawrence and Schmitt, 1993).

Face validity, or the extent to which the test appears to measure the construct for which it is intended, increased with the content changes because the items are now more reminiscent of ideas and methods of evaluation taught in the classroom. Also, the content changes provide more content validity based on reasoning abilities from cognitive and learning theory. This is exemplified by the SPR items in the Math section and more emphasis on contextual reading abilities in the Verbal section.

Dorans and Feigenbaum (1993) discuss the issues and research regarding the equatability of the new test to the old one which could be considered a form of concurrent validity. This research is on the PSAT/NMSQT data from the field studies done in 1992, but it is assumed that the findings generalize to the SAT I because of the relationship between the two tests. While two tests can always be scaled the same, they are not truly equatable unless they measure the same construct. The method used to assess equatability was to determine if the equating process affected all subgroups the same. Dorans and Feigenbaum compared equated scores for subgroups converted using the total group and then only the subgroups. The Black subgroup was affected by the equating process. The Mathematics test mean based on total group conversion was 367 as compared to 371 based on the Black subgroup only conversion. This affected 13.1% of Black examinees by unrounded scaled score difference greater than five. For the Verbal section, the Male and Female subgroups were affected by the equating process, to the extent that about 66% of males and about 70% of females were affected. The mean for Males based on total group

conversion was 392 with subgroup conversion resulting in a mean of 400. The mean for Females based on the total group was 395, as opposed to 388 for the subgroup only conversion. This indicates that the equatability of the old and new tests is questionable. Dorans and Feigenbaum call for more research on actual test administrations.

Brennan (1993) voiced a concern about equating based on the APA, AERA and NCME (1985) Standards for Educational and Psychological Testing which state that if the statistical specifications or the content of a test is changed, the old and new scores should not be equated. It is quite clear that both the statistical specifications and the content of the SAT were changed. Also, there is evidence that the two tests measure different constructs, even by design.

Since the most common use of the SAT score is in college admission, it could be argued that predictive validity is the most important validity information. More actual administrations of the test need to occur and the test takers need to be admitted to and attend college before definitive studies can be done. Morgan (1994) discussed the predictive validity of the old scale as compared to the new recentered scale. He performed an in-depth study of over 39,000 students from 45 colleges and universities taking the SAT in 1985. The colleges and universities provide a good cross-section of the population by public and private school, geographic area, gender, race/ethnicity, and SAT ability. He studied the effects of scaling the students raw score to either the old scale or the new recentered scale on the predictive value for freshman GPA and course performance. He found that recentering the scale improved the predictive validity of the SAT on average and across most subgroups. For example, when considering freshman GPA for all students, the Verbal validity coefficient rose from .485 to .492, while the Mathematics validity coefficient rose from .509 to .515. When combined with high school record, the predictive validity rose from .644 to .649 and gave an SAT increment, or the difference between the predictive value of high school record and the predictive value of high school record combined with SAT, from .066 to .070.

The technical aspects of recentering are presented in Dorans (1994a, and 1994b). Basically, the scores were recentered at the mid-point of the 200-800 scale range with a standard deviation of 110. This standard deviation was chosen to spread out the scores and allow for differences in test form difficulty. With a standard deviation of 100, a perfect raw score on an easy form should not be scaled to 800, but would be because the test taker performed at 100%. With a standard deviation of 110, several

scores scale to 800, allowing for inevitable differences between theoretical and practical test specifications (Dorans, 1994a).

Recentering affects parts of the scale differently. The recentering of the Verbal test was fairly straightforward because the 1990 Reference Group scores were nearly normally distributed. Thus, by recentering, Verbal scores were moved up the scale 70 to 80 points. The Mathematics test scores were not normally distributed, however, and the conversion was different at various parts of the scale. Briefly, for the score ranges of 200-240 and from the high 600's to the low 700's, the new scores are actually lower than the old. From the high 500's to the low 600's, the scores are virtually unchanged, while for the upper 700's and the range from 250 to 550, the scores are higher (Dorans, 1994a). Conversion charts give estimates of 'equivalent' old and new scores (College Entrance Examination Board, 1995a and 1995b).

Practical Considerations

With such a large scale testing program that affects so many lives, the technical considerations have to be weighed against the practical considerations. Educational Testing Service and the College Board took these into consideration by not only examining and changing the technical aspects of the test, but also the registration, and interpretation materials for both the test takers and test users. In both cases, these materials are excellent in explaining the changes and interpreting results. A widespread campaign was undertaken a year before changes appeared on the actual tests to educate the public, as well as test takers and test users. These pamphlets and booklets are still available and advertised in the primary resources for test users (College Entrance Examination Board, 1995).

It is important to note that through both the changes in content and recentering, great pains were taken to ensure that the new test was not easier than the old one, even though mean test scores 'rose.'

For the Admissions Officer

The admissions officer primarily uses test results to decide which students to consider for admission. Two issues that affect these uses are equatability of old and new test scores, and the reliability, or precision, of the test scores. Because field tests indicated that subgroups were affected differently by the equating process, more research on actual administrations of the new test needs to be done, as indicated by Dorans (1994b), and the implications communicated to the test users. During the

transition years when some students have old scores and some have new, issues of fairness in admission practices arise, especially in enrollment management situations where a certain score is crucial to be considered for admission. If the scores are not equatable, the decision is made differently for students with old and new test scores. This raises the related question of whether admissions criteria need to be changed to accommodate the differences in the test. Perfetto and Sanders (1995) performed a study on the implications of the changes in the SAT for Vanderbilt University admissions. They did not address the issue of score cutoffs for admission consideration, but did consider the issue of comparing students for admission. They note that the recentering does not change the ordering of students, only the location on the scale. Another issue Perfetto and Sanders addressed is using predictive models for freshman GPA. The new scores cannot be substituted into the old models in place of the old scores. A new model must be made using new scores. This cannot be definitively done until sufficient data are collected on several classes with new scores. For now, admissions officers have been provided with conversion tables from old to new test scores and vice versa. It is important to note that these tables provide only estimates of 'equivalent' scores.

The other issue in using test scores to compare students is the reliability, or precision, of the test. The most utilitarian aspect of reliability is the Standard Error of Measurement (SEM). The SEM provides the bounds for the range within which the true test score is likely to reside. For interpretation of test results, old and new SEM's should be approximately equal and the SEM should be approximately equal across all test scores which was acheived through controlling the statistical specifications of the test (Lawrence and Schmitt, 1993).

For the Student

As part of the focus group discussions during the development of the new test, attention was given to interpreting test results to students. Students wanted more practical, rather than technical, information. The most notable changes are the inclusion of a section to help evaluate the decision to take the test again, and displays of the number and type of items missed, along with the estimated percentile compared to college-bound seniors (Cook, 1995). These changes, especially with the input of the test takers and users themselves behind them, can only help.

A consequence of the equatability question suggests that students who took the old test and are not satisfied with their scores should try taking the new test.

Summary

The SAT, as a major factor in the lives of college-bound students, their parents, high school teachers and guidance counselors, and college admissions and placement officers, has a responsibility to society to uphold its high standards of technical merit. Over the past half century, the SAT test taking population has changed drastically, along with theory about cognition, learning, and psychometrics. The benefits of change must always be weighed against the associated problems. The SAT program reached a point at which the benefits of change outweighed the problems that change, associated with long use, practice and public relations, would engender. In fact, it could be said that change was not only justified, but sorely needed in order to adequately serve the test taking and using populations. The changes in the SAT program were researched, developed and implemented in a responsible manner and the resulting test represents a better measure of the abilities of today's test takers than did the old form.

Three points need to be addressed further regarding the reliability, equatability and validity of the SAT I. Changes to test specifications on the new test, especially in the Math section, occurred after the field studies. Research as to the effects of these specifications are needed. Cook (1995) referenced an unpublished report. The results of these studies would be helpful to the test using population. Also, preliminary research from field tests suggested that the changes to the test affected subgroups of the test taking population differently, most seriously males and females. This, coupled with the concern that if the statistical specifications or the content of a test is changed, the old and new scores should not be equated (APA, AERA and NCME, 1985), indicates a serious need for further research on actual administrations of the new test, and for that research, regardless of the outcome, to be communicated to test takers and users in a timely fashion. Each piece of research presented in this paper covered the effects of either the scale change or the content change. Research taking into account all facets of the change needs to be undertaken to determine the overall effect. Also, as the first year of SAT I test takers finish their first year of college, more studies of the predictive validity of the new SAT I for freshman GPA would be helpful to test users, though it may be several years before enough data are accumulated to make the results trustworthy.

References

Anastasi, A. (1988). Psychological Testing (6th Ed.). New York: Macmillan Publishing Company.

American Psychological Association, American Educational Research Association & National Council on Measurement in Education. (1985). Standards for Educational and Psychological Testing. American Psychological Association, American Educational Research Association & National Council on Measurement in Education.

Brennan, R. L. (1993). The new SAT I and PSAT/NMSQT: a discussion of some psychometric issues. In Educational Testing Service. (1994, March). Technical issues related to the introduction of the new SAT and PSAT/NMSQT (RM-94-10). Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1995). Bulletin for the SAT Program: 1995-96 Registration. Educational Testing Service.

College Entrance Examination Board. (1995a). Admission Officer's handbook for the SAT Program. Princeton, New Jersey: College Entrance Examination Board and Educational Testing Service.

College Entrance Examination Board. (1995b). Counselor's handbook for the SAT Program. Princeton, New Jersey: College Entrance Examination Board and Educational Testing Service.

College Entrance Examination Board. (1991). SAT. College Entrance Examination Board.

Cook, L. L. (1994, April). Recentering the SAT score scale: and overview and some policy considerations. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Cook, L. L. (1995, April). Lessons learned: implementing change in the SAT. Draft of a paper to be presented at the annual meeting of the National Council on Measurement on Education, San Francisco.

Dodd, B. G. (1993). Discussant comments on the NCME Symposium "Technical issues related to the introduction of the new SAT and PSAT/NMSQT." In Technical issues related to the introduction of the new SAT and PSAT/NMSQT (RM-94-10). Princeton, NJ: Educational Testing Service.

Dorans, N. J. & Feigenbaum, M. D. (1993, April). Equating issues engendered by changes to the new SAT I and PSAT/NMSQT. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Dorans, N. J. (1994, May). Choosing and evaluating a scale transformation: centering and realigning SAT score distributions. Revision of a paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Dorans, N. J. (1994, May). Effects of scale choice on score distributions: two views of subgroup performance on the SAT. Revision of a paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Feryok, N. J. & Wright, N. K. (1993, April). Design of the new SAT and PSAT/NMSQT Spring 1992 field trial. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Lawrence, I. M. & Schmitt, A. P. (1993, April). Setting statistical specifications for the new SAT and PSAT/NMSQT. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Lawrence, I. (1995). Estimating reliability for tests composed of item sets. Draft of a paper to be presented at the annual meeting of the National Council on Measurement on Education, San Francisco.

Morgan, R. (1994, April). Effects of scale choice on predictive validity. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Perfetto, G. & Sanders N. (1995). Recentering: what does it really mean for admission? College Entrance Examination Board.